Homework Assignment 3

(200 points)

Assigned Date: Tuesday, March 7, 2013

Due Date: ~~4:00PM Tuesday, March 26 2013~~
Extended to 4:00 PM Thursday, March 28 2013

Educational Goal

Apply WEKA to understand and analyze the flooding prediction problem with the aid of flooding data visualization; Perform interdisciplinary teamwork between CS students and EEOS students.

Requirements

- **Flooding data**

  http://www.cs.umb.edu/~yangmu/dataset/data_precipitation.zip(~50MB)

  The zip file includes the raw data of:

  1. The map files are:

  sampleLocations.shp
  worldmap.shp
  states.shp

  2. The atmospheric variables (files) are:

  1000hPa geopotential height (Z1000.csv)
  500hP geopotential height (Z500.csv)
  300hPa geopotential height (Z300.csv)
  850hPa temperature (T850.csv)
  850hPa zonal and meridional wind (U850.csv and V850.csv)
  300hPa zonal and meridional wind (U300.csv and V300.csv)
  Precipitable water (PW.csv)
  Geopotential height is in meters, temperature in Kelvin, wind in meters per second and precipitable water in mm.

  The atmospheric data is from January 1, 2010 to December 31, 2010. Each atmospheric variable is stored as a 2-D matrix with each row being a daily average value and each column being a point in space. The columns are associated with the "ID" field in the map file of "samplelocations.shp". Each file represents 5,328 points in space between the equator and North Pole (37 latitudes and 144 longitudes). For example, the data in column 4 row 6 of the

Z1000.csv file represents the 1000hPa geopotential height in January 6, 2010 at the location marked by the point with "ID"=4 in the samplelocations.shp file.

3. Extreme Precipitation Cluster: There are frequent heavy precipitation events from early July to mid-August in the State of Iowa. Severe thunderstorm activity during August 8–11, 2010 in central and southeast Iowa resulted in major flooding from August 11–16, 2010. Two related "Blocking" events have been observed during July and August. Particularly, the extreme precipitation cluster we are looking for begins at July 4th, 2010.

   Description of Extreme Precipitation Cluster data:

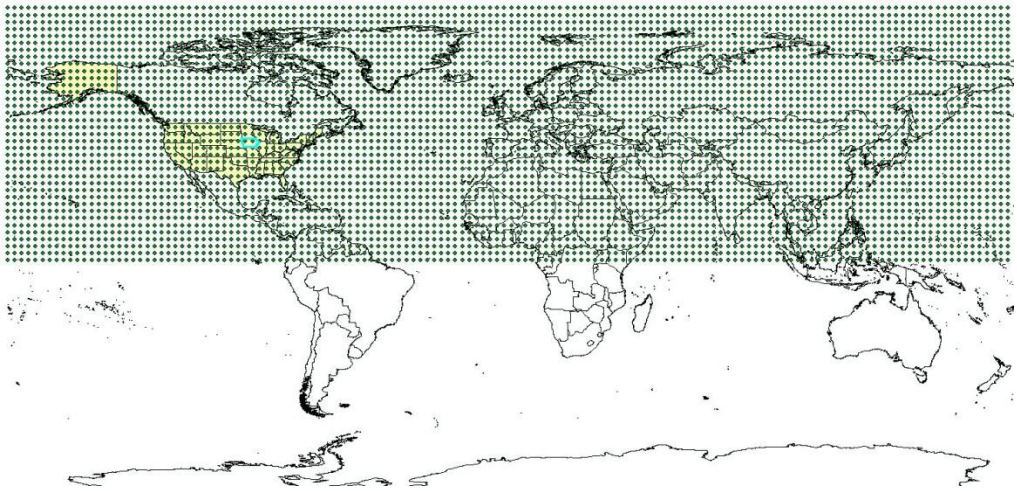
   Column 1: Spatial average precipitation (22 stations covering Iowa) for each day (inch)
   Column 2: Spatial standard deviation of daily precipitation at the 22 stations for each day
   Column 3: Day of the month
   Column 4: Julian day (1-365)
   Column 5: Month
   Column 6: Year

- **Goal:**

  Understand the data using visualization.

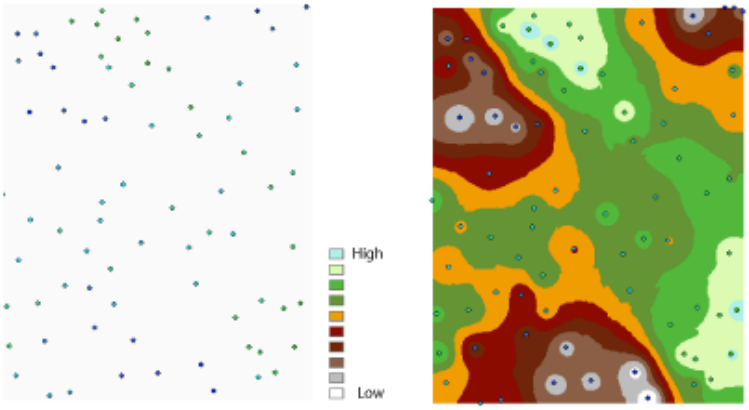
- **Tasks**

  1. Visualize the data points in a global map. For example:



  2. Visualize each factor (geopotential height, temperature, etc.) on the map using interpolation analysis in a global view. For a start day of an extreme precipitation cluster, visualize the result at least 10 days ahead it.

  For example: suppose 5/30/2011 is a start day of a precipitation cluster, visualize the data from 5/20/2011 to 5/29/2011. For each factor, extract the data for a day in that range, and visualize it on the map using interpolation analysis. A sample of interpolation result is:



3. Analyze your results and state your idea on how to find the most useful pattern resulting to Precipitation Clusters using visualization. Describe your idea and state why your idea can work (show reasons using visualized results) or does not work (show reasons using visualized results).

Tips: A pattern is a constrained factor. For example (the example may not be true in reality), low temperature (constraint of value) happened 5 days ago (constraint of time) at New York state (constraint of location) resulting a precipitation cluster in Iowa.

4. Plot the precipitation value (State of Iowa) versus the time change, noted as a curve c.

5. Plot the each factor value versus the time change, noted as curves f_1, f_2,..., f_n, given some places close to Iowa. Please pick those spatial points based on your observation.

6. Let's assume that curve c can be estimated by f_1, f_2,..., f_n using a linear combination. For flooding forecasting, we need to estimate c ahead of t days. Therefore, the problem becomes to find the best weight vector a, which has the form of $\hat{c}(x) = \sum_i a_i f_i(x - t)$, such that $error = \sum_x (\hat{c}(x) - c(x))^2$ is minimized (hint: this is a linear regression problem). Also try different t and plot a figure with $error$ versus t, where $t \geq 5$. Then study the results based on which t yields the minimum error.

7. Combine your conclusions from Tasks 3 to 6 and make another in-depth investigation cycle. Analyze whether results from Tasks 3 and 6 are confirmed. If not, try to do it in the other way. For example, check whether patterns observed in Task 3 can be observed and confirmed in the result of Task 6 (for example, has a relatively larger weight coefficient); check whether factors corresponding to high weight coefficients obtained in Task 6 imply any visible pattern using ArcGIS map visualization.

Submission Requirements

1. Write a brief report that records how your investigations proceeded and answering the above 6 questions.
2. Attach the homework cover page to your report
3. Submit the softcopy of your report and code via UMassOnline. Zero points for late submission.
4. Turn in the paper copy of your report in class. Paper copy should be bound firmly together as one pack (for example, staple, but not limited to, at the left corner). 5 points will be deducted

for unbounded homework.

5. No hard copies or soft copies results in 0 points.